

Class 17 Vaccination Mini Project

Pierce Ford (PID: A59010464)

11/24/2021

Background

In this before Thanksgiving class when many of our class mates are traveling let's have a look at COVID-19 vaccination rates around the state.

We get vaccination rate data from CA.gov here: <https://data.ca.gov/dataset/covid-19-vaccine-progress-dashboard-data-by-zip-code>

Import Data

```
vax <- read.csv("covid19vaccinesbyzipcode_test.csv")
head(vax)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction      county
## 1 2021-01-05                92395          San Bernardino San Bernardino
## 2 2021-01-05                93206                Kern      Kern
## 3 2021-01-05                91006          Los Angeles  Los Angeles
## 4 2021-01-05                91901          San Diego    San Diego
## 5 2021-01-05                92230          Riverside    Riverside
## 6 2021-01-05                92662            Orange      Orange
##   vaccine_equity_metric_quartile      vem_source
## 1                             1 Healthy Places Index Score
## 2                             1 Healthy Places Index Score
## 3                             3 Healthy Places Index Score
## 4                             3 Healthy Places Index Score
## 5                             1 Healthy Places Index Score
## 6                             4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                35915.3                40888                NA
## 2                 1237.5                 1521                NA
## 3                28742.7                31347                19
## 4                15549.8                16905                12
## 5                 2320.2                 2526                NA
## 6                 2349.5                 2397                NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                        NA                        NA
## 2                        NA                        NA
## 3                       873                0.000606
## 4                       271                0.000710
```

```
## 5 NA NA
## 6 NA NA
## percent_of_population_partially_vaccinated
## 1 NA
## 2 NA
## 3 0.027850
## 4 0.016031
## 5 NA
## 6 NA
## percent_of_population_with_1_plus_dose
## 1 NA
## 2 NA
## 3 0.028456
## 4 0.016741
## 5 NA
## 6 NA
## redacted
## 1 Information redacted in accordance with CA state privacy requirements
## 2 Information redacted in accordance with CA state privacy requirements
## 3 No
## 4 No
## 5 Information redacted in accordance with CA state privacy requirements
## 6 Information redacted in accordance with CA state privacy requirements
```

Q1. How many entries do we have?

```
nrow(vax)
```

```
## [1] 82908
```

We can use the **skimr** package and the **skim()** function to get a quick overview of the structure of this dataset.

```
skimr::skim(vax)
```

Table 1: Data summary

Name	vax
Number of rows	82908
Number of columns	14
Column type frequency:	
character	5
numeric	9
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
as_of_date	0	1	10	10	0	47	0
local_health_jurisdiction	0	1	0	15	235	62	0
county	0	1	0	15	235	59	0
vem_source	0	1	15	26	0	3	0
redacted	0	1	2	69	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
zip_code_tabulation_area	0	1.00	93665.111817.39	90001	92257.7593658.5095380.5097635.0					
vaccine_equity_metric_quartile	0	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.0418993.94	0	1346.95	13685.1031756.1288556.7				
age5_plus_population	0	1.00	20875.2421106.04	0	1460.50	15364.0034877.00101902.0				
persons_fully_vaccinated	8355	0.90	9585.35	11609.12	11	516.00	4210.00	16095.0071219.0		
persons_partially_vaccinated	8355	0.90	1894.87	2105.55	11	198.00	1269.00	2880.00	20159.0	
percent_of_population_fully_vaccinated	8355	0.90	0.43	0.27	0	0.20	0.44	0.63	1.0	
percent_of_population_partially_vaccinated	8355	0.90	0.10	0.10	0	0.06	0.07	0.11	1.0	
percent_of_population_with_8355plus_dose	8355	0.90	0.51	0.26	0	0.31	0.53	0.71	1.0	

Notice that one of these columns is a date column. Working with time and dates gets annoying quickly. We can use the **lubridate** package to make this easy.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##   date, intersect, setdiff, union
```

```
#example lubridate function
today()
```

```
## [1] "2021-11-24"
```

Q2. How many days since the first entry in the dataset?

```
vax$as_of_date[1]
```

```
## [1] "2021-01-05"
```

```
#Change this to a date object and do math with today
today() - ymd(vax$as_of_date[1])
```

```
## Time difference of 323 days
```

```
#This works, so let's update the whole column  
vax$as_of_date <- ymd(vax$as_of_date)
```

```
#Test that overwrite worked  
today() - vax$as_of_date[1]
```

```
## Time difference of 323 days
```

Q3. When was the dataset last updated? What is the last date in this dataset and how many days have passed since then?

Q4. How many days does the dataset span?

```
#Last day  
last_day <- vax$as_of_date[nrow(vax)]  
last_day
```

```
## [1] "2021-11-23"
```

```
#Days since last day  
today() - last_day
```

```
## Time difference of 1 days
```

```
#Span of data  
last_day - vax$as_of_date[1]
```

```
## Time difference of 322 days
```

Q5. How many different ZIP code areas are in this dataset?

```
#Saving the zip codes as factors might be useful in describing the data "by ZIP code" later  
zip_factors <- as.factor(vax$zip_code_tabulation_area)  
#Print how many different "levels" (ZIP codes) there are  
length(levels(zip_factors))
```

```
## [1] 1764
```

To work with ZIP codes, we can use **zipcodeR**

```
library(zipcodeR)  
  
#example of using zipcodeR to pull up real world info about the ZIP codes  
reverse_zipcode(c('92037', '92109'))
```

```
## # A tibble: 2 x 24  
##   zipcode zipcode_type major_city post_office_city common_city_list county state  
##   <chr>    <chr>         <chr>      <chr>                                <blob> <chr>  <chr>
```

```
## 1 92037 Standard La Jolla La Jolla, CA <raw 20 B> San D~ CA
## 2 92109 Standard San Diego San Diego, CA <raw 21 B> San D~ CA
## # ... with 17 more variables: lat <dbl>, lng <dbl>, timezone <chr>,
## # radius_in_miles <dbl>, area_code_list <blob>, population <int>,
## # population_density <dbl>, land_area_in_sqmi <dbl>,
## # water_area_in_sqmi <dbl>, housing_units <int>,
## # occupied_housing_units <int>, median_home_value <int>,
## # median_household_income <int>, bounds_west <dbl>, bounds_east <dbl>,
## # bounds_north <dbl>, bounds_south <dbl>
```

Focus in on San Diego County

Changing track from ZIP codes for a moment. We want to subset the full vax data down to just San Diego County.

```
#Subset and check with base R
vax_san_diego_base <- vax[vax$county == "San Diego",]
nrow(vax_san_diego_base)
```

```
## [1] 5029
```

```
head(vax_san_diego_base)
```

```
## as_of_date zip_code_tabulation_area local_health_jurisdiction county
## 4 2021-01-05 91901 San Diego San Diego
## 14 2021-01-05 91902 San Diego San Diego
## 21 2021-01-05 92011 San Diego San Diego
## 22 2021-01-05 92055 San Diego San Diego
## 25 2021-01-05 92067 San Diego San Diego
## 33 2021-01-05 92081 San Diego San Diego
## vaccine_equity_metric_quartile vem_source
## 4 3 Healthy Places Index Score
## 14 4 Healthy Places Index Score
## 21 4 Healthy Places Index Score
## 22 3 CDPH-Derived ZCTA Score
## 25 4 Healthy Places Index Score
## 33 2 Healthy Places Index Score
## age12_plus_population age5_plus_population persons_fully_vaccinated
## 4 15549.8 16905 12
## 14 16620.7 18026 22
## 21 20503.6 23247 NA
## 22 11548.0 11654 NA
## 25 6973.9 7480 11
## 33 25558.0 27632 14
## persons_partially_vaccinated percent_of_population_fully_vaccinated
## 4 271 0.000710
## 14 374 0.001220
## 21 NA NA
## 22 NA NA
## 25 241 0.001471
## 33 346 0.000507
## percent_of_population_partially_vaccinated
```

```
## 4 0.016031
## 14 0.020748
## 21 NA
## 22 NA
## 25 0.032219
## 33 0.012522
## percent_of_population_with_1_plus_dose
## 4 0.016741
## 14 0.021968
## 21 NA
## 22 NA
## 25 0.033690
## 33 0.013029
## redacted
## 4 No
## 14 No
## 21 Information redacted in accordance with CA state privacy requirements
## 22 Information redacted in accordance with CA state privacy requirements
## 25 No
## 33 No
```

Subsetting can get tedious and complicated when we have multiple criteria, so we can use the **dplyr** package.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
## filter, lag

## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union
```

We will use the `filter()` function to do our subsetting since it is more robust. Focus in on San Diego County as before.

```
#Get San Diego entries and check
sd <- filter(vax, county=="San Diego")
nrow(sd)
```

```
## [1] 5029
```

More complicated subsetting:

```
#Filter by county AND population and check
sd.20 <- filter(vax, county=="San Diego",
                age5_plus_population > 20000)
nrow(sd.20)
```

```
## [1] 3055
```

Q6. What is the average vaccination rate of San Diego County as of yesterday?

```
#Filter and check
sd.now <- filter(vax, county=="San Diego",
                 as_of_date=="2021-11-23")
head(sd.now)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction   county
## 1 2021-11-23                92120          San Diego San Diego
## 2 2021-11-23                91962          San Diego San Diego
## 3 2021-11-23                92155          San Diego San Diego
## 4 2021-11-23                92147          San Diego San Diego
## 5 2021-11-23                91913          San Diego San Diego
## 6 2021-11-23                92114          San Diego San Diego
##   vaccine_equity_metric_quartile          vem_source
## 1                        4 Healthy Places Index Score
## 2                        3 Healthy Places Index Score
## 3                       NA          No VEM Assigned
## 4                       NA          No VEM Assigned
## 5                        3 Healthy Places Index Score
## 6                        2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                26372.9                28414                21234
## 2                 1758.7                 2020                 948
## 3                  456.0                  456                  70
## 4                  518.0                  518                  NA
## 5                43514.7                50461                37974
## 6                59050.7                64945                43708
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                        3198                        0.747308
## 2                         126                        0.469307
## 3                          20                        0.153509
## 4                         NA                        NA
## 5                       6690                        0.752542
## 6                       6261                        0.673000
##   percent_of_population_partially_vaccinated
## 1                        0.112550
## 2                        0.062376
## 3                        0.043860
## 4                         NA
## 5                        0.132578
## 6                        0.096405
##   percent_of_population_with_1_plus_dose
## 1                        0.859858
## 2                        0.531683
## 3                        0.197369
## 4                         NA
## 5                        0.885120
## 6                        0.769405
##
##                                     redacted
## 1                                     No
```

```
## 2 No
## 3 No
## 4 Information redacted in accordance with CA state privacy requirements
## 5 No
## 6 No
```

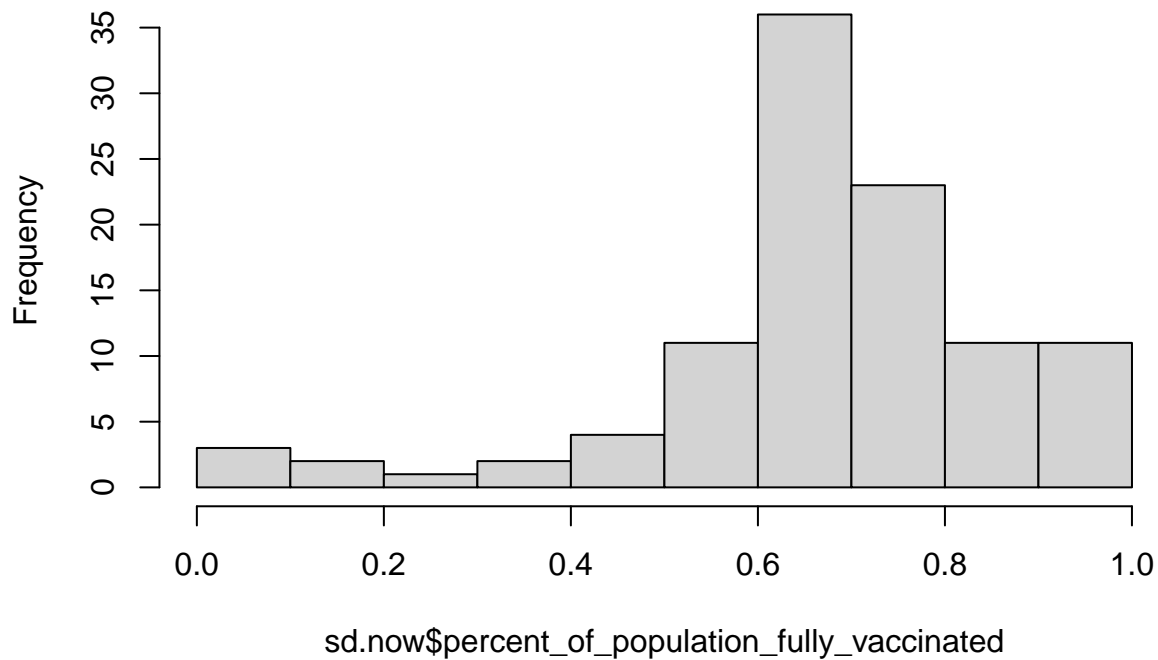
```
#Summary of vaccination percent, including mean
summary(sd.now$percent_of_population_fully_vaccinated)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
## 0.01017 0.61301 0.67965 0.67400 0.76932 1.00000      3
```

Q7. Make a histogram of these values.

```
#Base R
hist(sd.now$percent_of_population_fully_vaccinated)
```

Histogram of sd.now\$percent_of_population_fully_vaccinated



However, this method skews for small counties (strong per-person effect on percentages).

Q8. What is the population of the 92037 ZIP code (UCSD) area?

Q9. What is the average vaccination value for this UCSD/La Jolla ZIP code area?


```
#Filter and view
```

```
lj <- filter(sd.now, zip_code_tabulation_area=="92037")
lj
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction   county
## 1 2021-11-23                92037                San Diego San Diego
##   vaccine_equity_metric_quartile                vem_source
## 1                        4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                33675.6                36144                33115
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                        7660                        0.916196
##   percent_of_population_partially_vaccinated
## 1                        0.21193
##   percent_of_population_with_1_plus_dose redacted
## 1                        1           No
```

```
#Population and vaccination rate
```

```
lj$age5_plus_population
```

```
## [1] 36144
```

```
lj$percent_of_population_fully_vaccinated
```

```
## [1] 0.916196
```

Q10. What about my Thanksgiving ZIP code?

```
home.vax <- filter(vax, zip_code_tabulation_area=="92882",
                  as_of_date=="2021-11-23")
```

```
#Percent Vaxxed
```

```
home.vax$age5_plus_population
```

```
## [1] 68310
```

```
home.vax$percent_of_population_fully_vaccinated
```

```
## [1] 0.564632
```

Time Course for Here and Home

Start with time course for 92037.

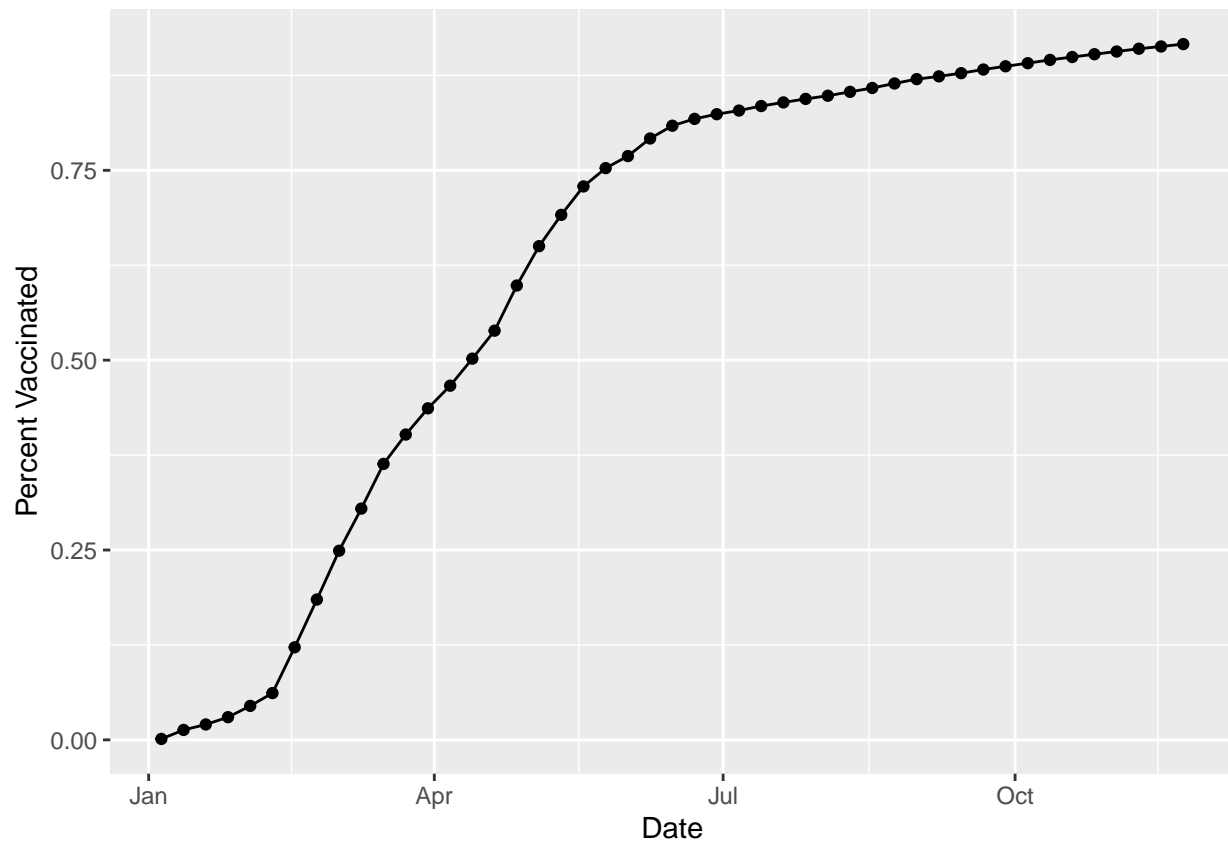
```
#load ggplot
```

```
library(ggplot2)
```

```
#filter for here (92037)
```

```
here.by_date <- filter(vax, zip_code_tabulation_area=="92037")
```

```
#plot here
ggplot(here.by_date) +
  aes(x=as_of_date, y=percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  labs(x="Date", y="Percent Vaccinated")
```



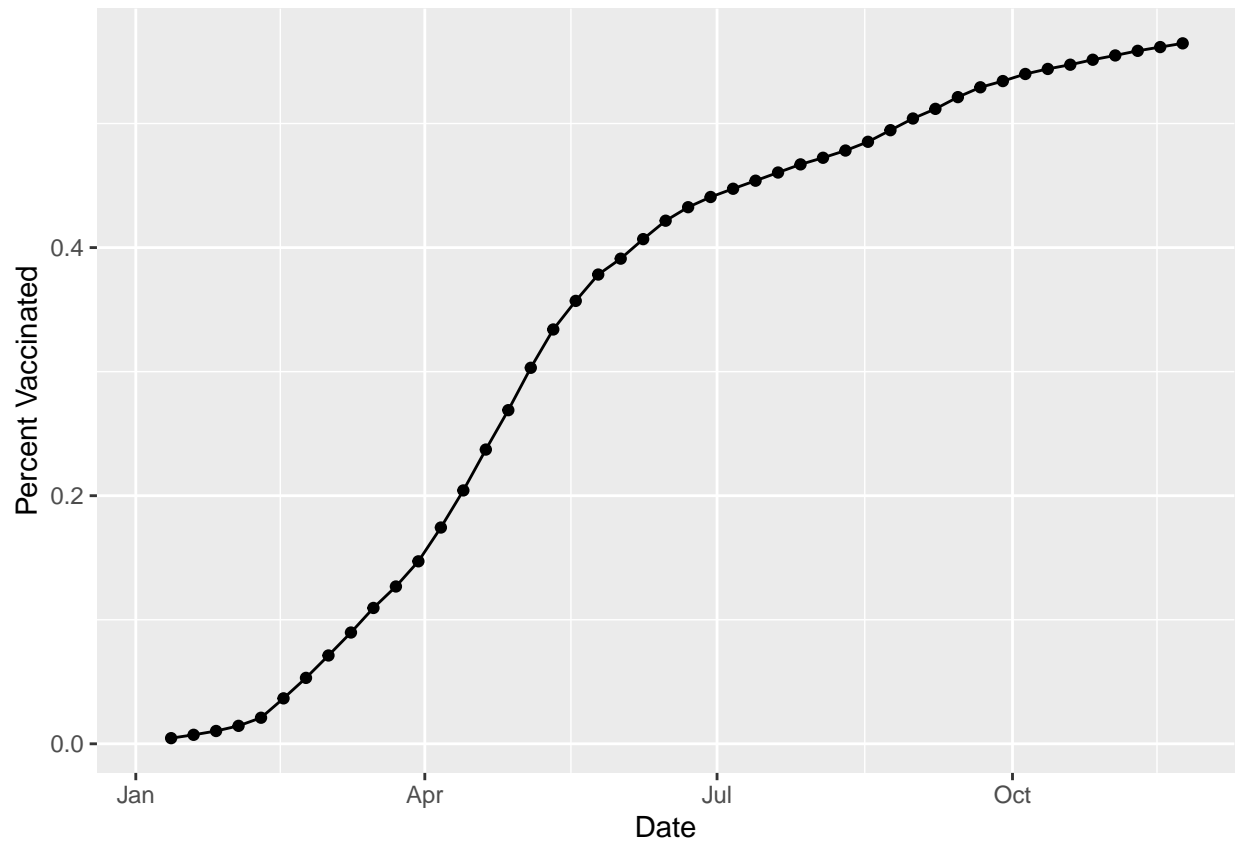
Repeat for Home (92882).

```
#filter for home (92882)
home.by_date <- filter(vax, zip_code_tabulation_area=="92882")

#plot here
ggplot(home.by_date) +
  aes(x=as_of_date, y=percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  labs(x="Date", y="Percent Vaccinated")
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



Let's compare across all of San Diego county.

Q11. How many ZIP codes pass this filter?

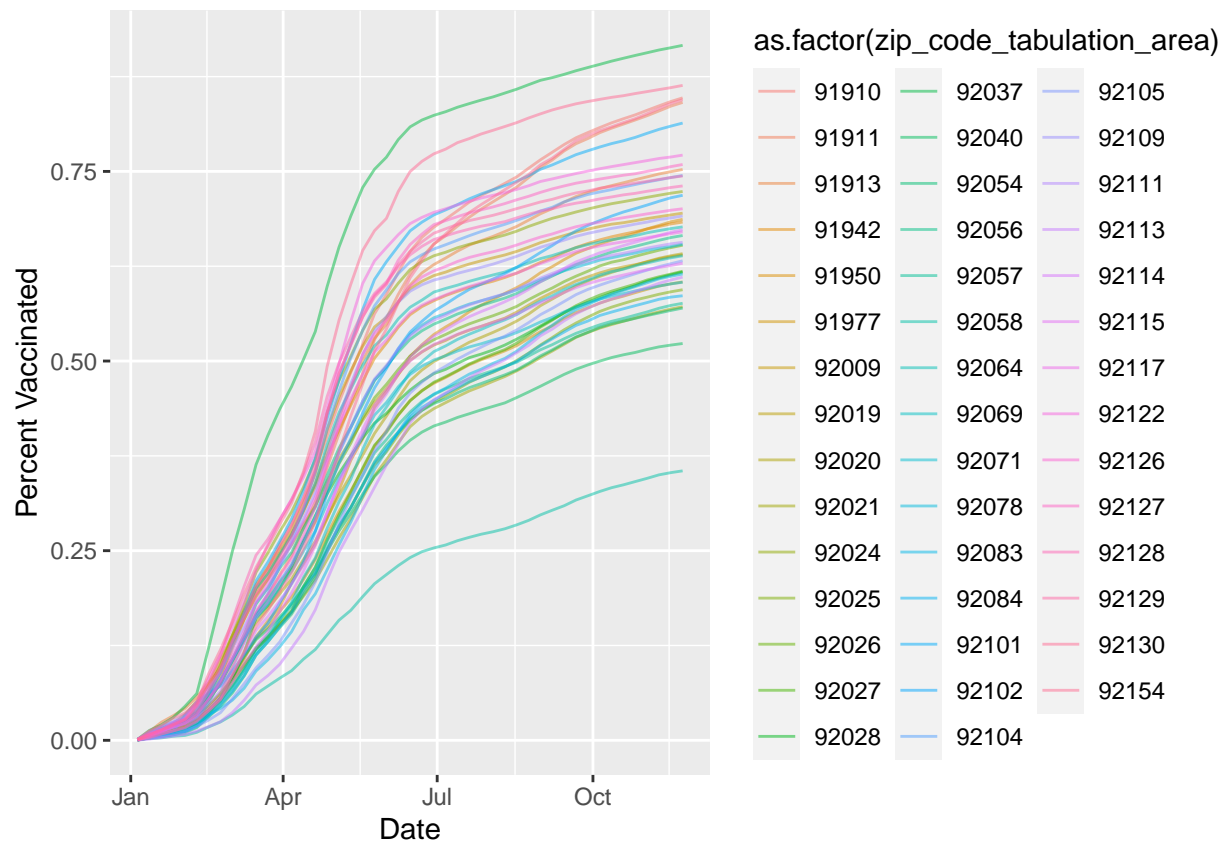
```
#filter for large zip codes
sd.large <- filter(vax, county=="San Diego",
                   age5_plus_population >= 36144)

#how many ZIPs
length(levels(as.factor(sd.large$zip_code_tabulation_area)))
```

```
## [1] 44
```

```
#plot sd.large
ggplot(sd.large) +
  aes(x=as_of_date, y=percent_of_population_fully_vaccinated, col=as.factor(zip_code_tabulation_area),
      geom_line(alpha=0.5) +
      labs(x="Date", y="Percent Vaccinated"))
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



Repeat for all of California.

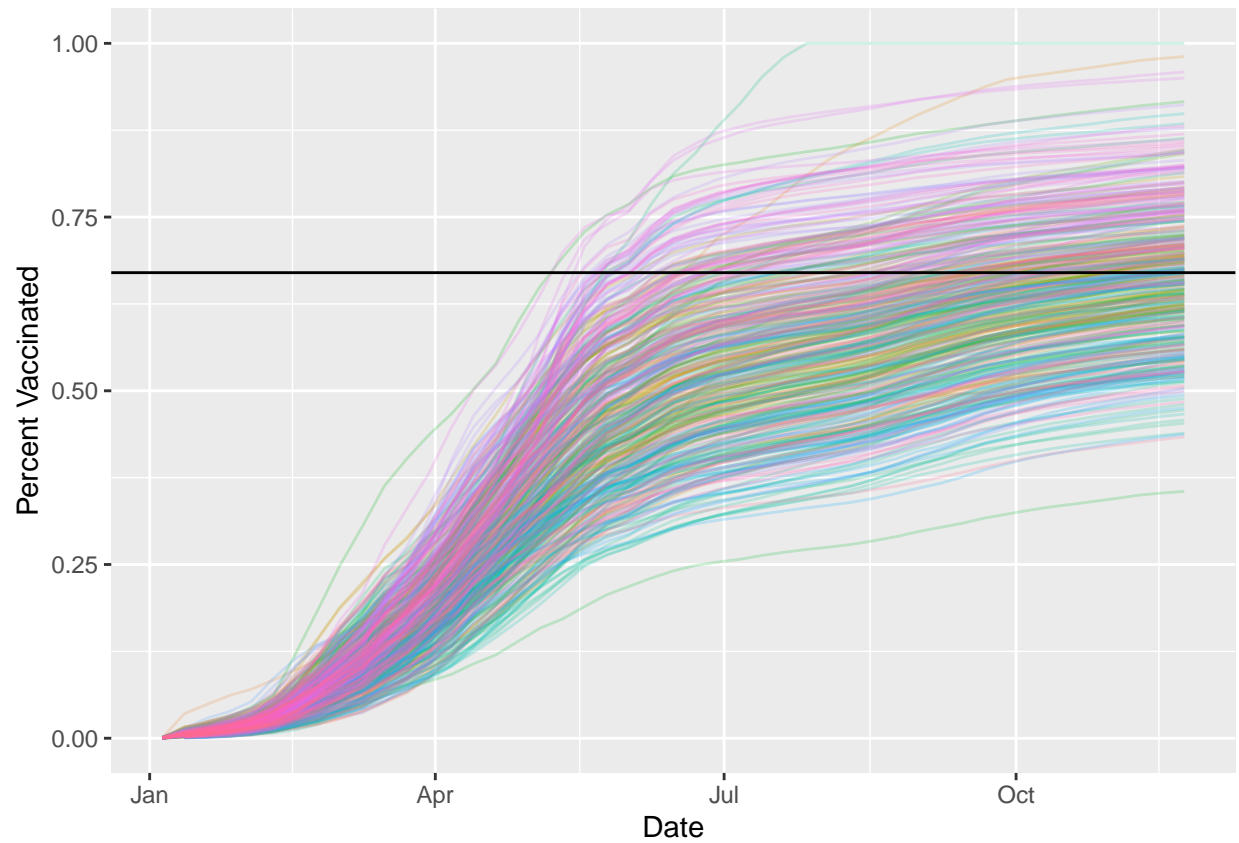
```
#filter for large zip codes
ca.large <- filter(vax, age5_plus_population >= 36144)

#how many ZIPs
length(levels(as.factor(ca.large$zip_code_tabulation_area)))
```

```
## [1] 412
```

```
#plot ca.large
ggplot(ca.large) +
  aes(x=as_of_date, y=percent_of_population_fully_vaccinated, col=as.factor(zip_code_tabulation_area),
    geom_line(alpha=0.2) +
    labs(x="Date", y="Percent Vaccinated") +
    theme(legend.position = "none") +
    geom_hline(yintercept = 0.67)
```

```
## Warning: Removed 176 row(s) containing missing values (geom_path).
```



What is the mean across the state for these large pop ZIP codes?

```
#Filter for yesterday
ca.large.now <- filter(ca.large, as_of_date=="2021-11-23")

#Summary of vaccination percent, including mean
summary(ca.large.now$percent_of_population_fully_vaccinated)
```

##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	0.3552	0.5939	0.6698	0.6678	0.7350	1.0000